Crystal structure of an intracellular protease from *Pyrococcus horikoshii* at 2-Å resolution

Xinlin Du*†, In-Geol Choi*, Rosalind Kim‡, Weiru Wang*, Jaru Jancarik*, Hisao Yokota‡, and Sung-Hou Kim*‡§

*Department of Chemistry, University of California, Berkeley, CA 94720; and ‡Structural Biology Department, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Contributed by Sung-Hou Kim, October 23, 2000

The intracellular protease from *Pyrococcus horikoshii* (PH1704) and PfpI from *Pyrococcus furiosus* are members of a class of intracellular proteases that have no sequence homology to any other known protease family. We report the crystal structure of PH1704 at 2.0-Å resolution. The protease is tentatively identified as a cysteine protease based on the presence of cysteine (residue 100) in a nucleophile elbow motif. In the crystal, PH1704 forms a hexameric ring structure, and the active sites are formed at the interfaces between three pairs of monomers.

The *Pyrococcus horikoshii* 1704 gene product (PH1704) has extensive sequence homology (90% identity) to a *Pyrococcus furiosus* intracellular protease, PfpI. PfpI is characterized by its proteolytic activity and remarkable stability (1, 2). Although PfpI has no detectable sequence homology to any other member of a known protease family, antibodies raised against it crossreact with bovine pituitary proteasome (3). A PSI-BLAST (4) search of databases revealed that PH1704 has homologs in most organisms, although homologs with more than 30% sequence identity are found only in archaea and bacteria. A sequence alignment of PH1704 and several of its homologous proteins is shown in Fig. 1. Only PfpI, among the proteins listed in Fig. 1, has been characterized biochemically, and none have been studied structurally. We have overexpressed the PH1704 gene in *Escherichia coli*, purified and expressed the protein, and determined its three-dimensional (3D) structure by x-ray crystallography. Here, we report preliminary biochemical data on PH1704 and its 3D crystal structure at 2.0-Å resolution.

Methods

Bacterial Expression and Protein Purification. The cloning of PH1704 from *P. horikoshii* genomic DNA was carried out according to the "sticky-end PCR" method (5). Two pairs of primers were used to produce two PCR products, which were mixed, heated, and cooled to produce a DNA fragment with *Nde*I *and Bam*HI sticky ends, which was in turn inserted into pET21a (6). A selenomethionine derivative of the protein was expressed in a methionine auxotroph, *E. coli* strain B834 (DE3)/pSJS1244 (7, 8) grown in M9 medium supplied with selenomethionine. In the purification process, the cell lysate was subjected to heating (80°C for 30 min), anion exchange (HiTrap-Q), and size-exclusion column chromatography (Superdex 75). To avoid potential oxidation of the protein, 10 mM DTT was used in all buffers. The yield of protein was typically 15 mg of pure protein/liter of culture. Initial crystallization conditions were screened by using the sparse matrix method (ref. 9; Hampton Research, Laguna Niguel, CA) at room temperature. One microliter of 20 mg/ml PH1704 in 20 mM Tris·HCl, pH 7.5, and 1 mM EDTA was mixed with 1 μl of 0.1 M trisodium citrate dihydrate, pH 5.6, 0.2 M potassium tartrate tetrahydrate, 2.0 M ammonium sulfate, and equilibrated with 0.5 ml of the same solution in the reservoir by the vapor diffusion sitting drop method. Diffraction-quality crystals were obtained 2 days after setup.

Data Collection and Reduction. X-ray diffraction data sets were collected at three wavelengths at the Macromolecular Crystallography Facility beamline 5.0.2 at the Advanced Light Source at Lawrence Berkeley National Laboratory. The crystal was soaked in a drop of mother liquor with 30% glycerol (about 50 μl) for about 40 s before being flash-frozen in liquid nitrogen and exposed to x-ray. All data sets were processed with DENZO (10) and reduced with SCALEPACK (10) and programs in the CCP4 package (11). The statistics of the data collection and reduction are shown in Table 1 and Table 2, respectively.

Model Building and Refinement. The program SOLVE (12) was used to locate the selenium sites in the crystal and to calculate initial phases. The initial multiwavelength anomalous dispersion phases (13) were further improved by solvent flattening and histogram matching with the DM program in the CCP4 package (11). The map calculated by using the improved phases was of excellent quality. Three models were built by using the O program (14) and refined by using CNS (15). No noncrystallographic symmetry (NCS) constraints or restraints were applied in the last few cycles of refinement. The refinement statistics are shown in Table 2. The atomic coordinates and structure factors have been deposited into the Brookhaven Protein Data Bank with the accession number 1G2I.

Results

Biochemical Characterization of PH1704. The ORF of gene PH1704 codes for a polypeptide chain of 166 amino acids. The purified protein appears as four bands on an SDS/PAGE gel, with apparent molecular masses of 18, 40, 90, and 200 kDa (Fig. 2*A*). Prolonged heating in the presence of SDS increases the fraction of monomeric protein (18 kDa; Fig. 2*A*, lane 2). A gelatin-SDS/PAGE protease assay was carried out according to the procedure established for PfpI (1) to determine the proteolytic activity of the individual protein bands. The 200-kDa band showed the highest activity in the SDS/PAGE gelatin overlay assay (Fig. 2*B*), although it makes up only a small fraction of the total protein. Weaker activities were shown by the 90-kDa and 40-kDa bands (Fig. 2*B*). N-terminal sequencing of the blotted protein from the 200-kDa band identified it unambiguously as PH1704 (data not shown).

Abbreviations: PH1704, *Pyrococcus horikoshii* 1704 gene product; PfpI, *Pyrococcus furiosus* intracellular protease; NCS, noncrystallographic symmetry; rmsd, rms deviation.

Data deposition: The atomic coordinates have been deposited in the Protein Data Bank, www.rcsb.org (PDB ID code 1G2I).

†Present address: Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75235.

§To whom reprint requests should be addressed. E-mail address: shkim@cchem.berkeley.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.260503597. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.260503597

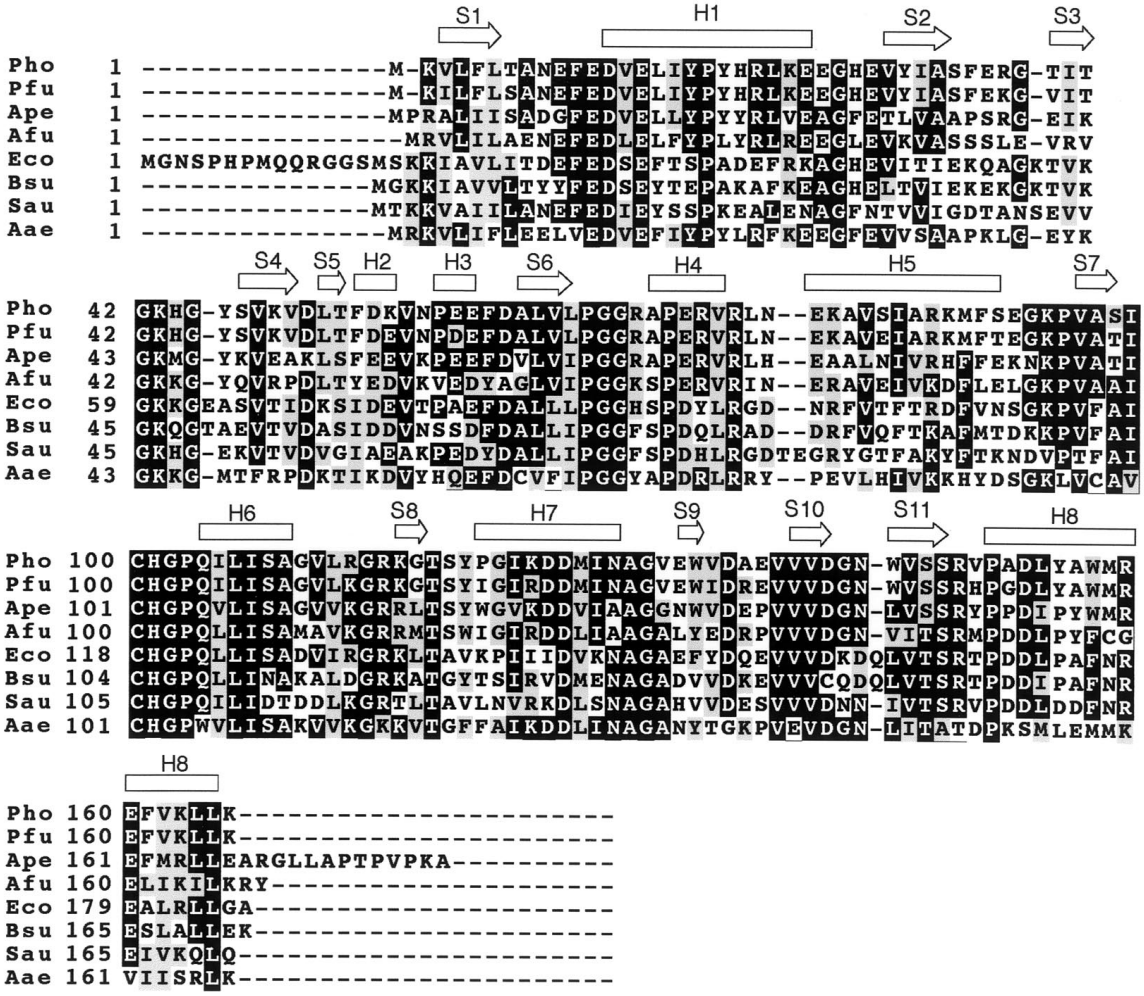

Fig. 1. Sequence alignment of PH1704 homologs. Abbreviations are: Pho, *P. horikoshii*; Pfu, *P. furiosus* (NCBI access ID, Q51732; percentage identity, 90%); Ape, *Aeopyrum pernix* (BAA79274.1, 61%); Afu, *Archaeoglobus fulgidus* (2649300, 53%); Eco, *E. coli* (P45470, 47%); Bsu, *Bacillus subtilis* (BAA23403, 43%); Sau, *Staphylococcus aureus* (Q53719, 39%); Aae, *Aquifex aeolicus* (2983230, 42%). Secondary structure analysis is based on the crystal structure of PH1704. ''H'' and ''S'' refer to helix and strand, respectively.

These results suggest that the most active form of the protein is a partially SDS-resistant, multimeric complex corresponding to the 200-kDa band.

Structure Determination. The PH1704 crystals belong to the space group $P4_12_12$ with cell dimensions of $a = b = 124.7$ Å, and $c = 129$ Å. There are three monomers in an asymmetric unit. Three models, A, B, and C for the three chains, were built by using the O program (14). The numbering of the residues is from 1 to 166, 201 to 366, and 401 to 566, for monomers B, A, and C, respectively.

Table 1. Data collection

Measurement	Value		
	Peak	Edge	Remote
Wavelength, Å	0.97938	0.9796	0.9686
Resolution, Å	2.0	2.0	2.0
No. of unique data	129,941	130,210	129,727
Redundancy	5.2	5.3	5.3
Overall completeness	98.9 (95.6)	98.9 (95.8)	98.9 (95.7)
R_{merge}, %	7.3 (34.6)	6.6 (31.8)	6.4 (35.3)
Average $I/\sigma(I)$	21.9 (4.9)	23.3 (5.2)	19.3 (3.7)

Space group $P4_12_12$. Cell: $a = b = 124.7$ Å, $c = 129.0$ Å. Each protein monomer has four selenomethionines. Values in parentheses are for highest-resolution shell, from 2.03 to 2.00 Å.

Monomer Structure. Each monomer consists of an α/β sandwich. A secondary structure analysis and ribbon diagram of the structure are shown in Fig. 1 and Fig. 3*A*, respectively. There are 11 β strands and eight helices as determined by the Database of Secondary Structure of Proteins (DSSP; ref. 16). The central β sheet consists of six β strands (S2, S1, S6, S7, S11, and S10). It is flanked by helices H8 and H1, and strands S3 and S4 on one side, and by helices H2, H3, H4, H5, H6, H7, and strands S9 and S8 on the other side.

The structure was compared with protein structures in the Protein Data Bank (PDB) by the DALI program (17). The two proteins with the highest Z scores are the noncatalytic domain of *E. coli* catalase HPII which has no known function (18), and the glutamine amidotransferase (GA) domain of GMP synthetase (19). Fig. 3*B* shows the Cα trace of GA aligned with that of PH1704. These two enzymes hydrolyze chemically related substrates, an amide bond in the case of GA and a

Table 2. Refinement statistics

Measurement	Value
Resolution, Å	20–2.0
No. of unique reflections	65,700
No. of parameters to fit	12,660
σ cutoff	None
R factor, %	18.4
R_{free}, %	20.0
rmsd bonds, Å	0.005
rmsd angles, °	1.2
rmsd NCS,* Å	0.12
Average *B* factor of protein	25.2
Average *B* factor of water	32.7
No. of water molecules	279
Residues in Ramachandran plot, %	
Most favored	90.6
Additional	8.7
Disallowed†	0.7

*Three models were built. The rmsd is an average of the rmsd values between AB, BC, and AC.

†The only residue in a disallowed main chain conformation is Cys-100. This is consistent with its proposed role as the nucleophile (see text).

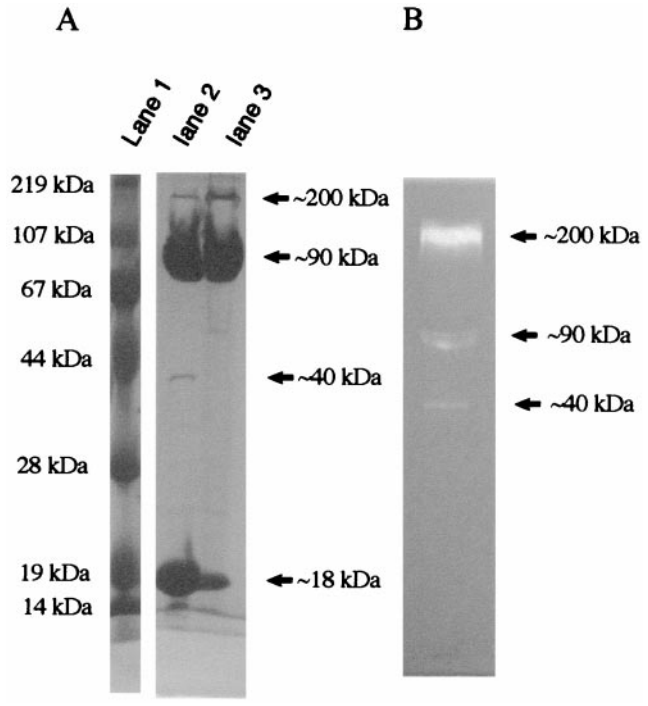

Fig. 2. Electrophoretic migration of PH1704 and gelatin protease assay. (*A*) 10% SDS-PAGE gel of purified PH1704. Lane 1, molecular markers; lane 2, PH 1704 boiled for 20 min in sample buffer; lane 3, PH1704 in sample buffer without boiling. (*B*) Gelatin overlay of 8% SDS-PAGE gel of purified PH1704. Proteolytic activity is manifested by cleared zones in the dark background. Molecular weights indicated are based on molecular markers in gelatin-SDS-PAGE gel.

peptide bond in the case of PH1704. In addition to the overall similarity in the folding topology between PH1704 and GA, there is a ''nucleophile elbow'' motif in both structures. The nucleophile elbow is a distinctive strand–nucleophile–helix motif that was first recognized in α/β hydrolases (20). The nucleophile of a catalytic triad in an α/β hydrolase, either a cysteine or a serine, resides in a sharp turn that connects the strand and the helix. Residues 96–109 (S7 and H6) of PH1704 form a nucleophile elbow-like motif (Fig. 3*C*). The Cα traces of this fragment can be aligned with the nucleophile elbow in amidotransferase with an rms deviation (rmsd) of 1.2 Å. The ϕ/ψ angles for the potential nucleophile Cys 100 of PH1704 fall in an unfavorable region in the Ramachandran plot; this is characteristic of the nucleophile in a nucleophile elbow. The sequence around Cys-100, S-I-C-H-G-P, is also consistent with the consensus sequence small-x-Nu-x-small-small for α/β hydrolases.

Quaternary Structure. Because only oligomeric forms of the protein showed activity, we looked closely into the quaternary structure. There are six kinds of intermolecular contacts in the crystal (Fig. 4*A*). The contacts between monomer B and C as well as that between A and B are symmetric and bury a significant amount of surface areas (1689 Å^2 and 1431 Å^2, respectively). The areas buried by the other four contacts are much smaller, and vary from 675 Å^2 to 348 Å^2.

In the case of the AB contact, there are a pair of salt bridges between Arg-22 and Glu-225 and the symmetry-related Arg-222 and Glu-25. Otherwise, the interactions are mainly hydrophobic. Residues that are buried in the interface include Tyr-18, Tyr-218, Val-14, Val-214, Leu-154, and Leu-354. Several other hydrophobic residues, including Tyr-46, Tyr-246, Ile-17, and Ile-217, are partially buried at the interface. In contrast, contact BC is primarily ionic in nature. There is hydrogen-bonding between His-101 and Glu-474, Ser-108 and Asp-525, and their symmetry-related counterparts, as well as electrostatic interactions between Arg-77 and Asp-526 and its symmetric pair Asp-126 and Arg-477. In addition, Ile-123, Ile-523, Ile-107, and Ile-527 are also buried in this contact. Both contacts, BC and AB, are likely to be biologically relevant instead of fortuitous crystal packing. In addition to the fact that these two buried surface areas are considerably larger than that of the other four intermolecular contacts, many of the residues involved in the ionic interactions and most of the hydrophobic residues involved in these two contacts are conserved among close homologs (Fig. 1).

Six monomers in the crystal are connected through the two major contacts described above to form a closed ring. The ribbon diagram and the surface of the hexamer are shown in Fig. 4 *A* and *B*, respectively. Monomer A, B, and C are crystallographically determined whereas D, E, and F are generated by a crystallographic twofold symmetry operation (around axis 1). Thus, the complex can be considered as a dimer of trimers, because there are two NCS twofold axes (axis 2 and 3) and one crystallographic twofold axis (axis 1) perpendicular to the threefold axis. Axis 2 relates trimer BCD to trimer AFE, and axis 3 relates trimer CDE to trimer BAF, with rmsd values of 0.42 and 0.42 Å, respectively. The complex can be considered also as a trimer of dimers because there is an NCS threefold axis, which relates dimer BC to DE, DE to FA, and FA to BC, with an rmsd of 0.63 Å.

Discussion

The activity of an intracellular protease must be strictly regulated to prevent cytoplasmic proteins from unwanted proteolytic degradation. This is ensured by cells in several different ways. Some proteases, like calpains and caspases, are highly specific (21, 22). Others are confined in endosomal and lysosomal compartments by a lipid membrane (cathepsins; refs. 23 and 24). ATP-dependent proteases, including proteosomes (25, 26), CLpP (27), Lon, and HslV (28), employ a

BIOCHEMISTRY

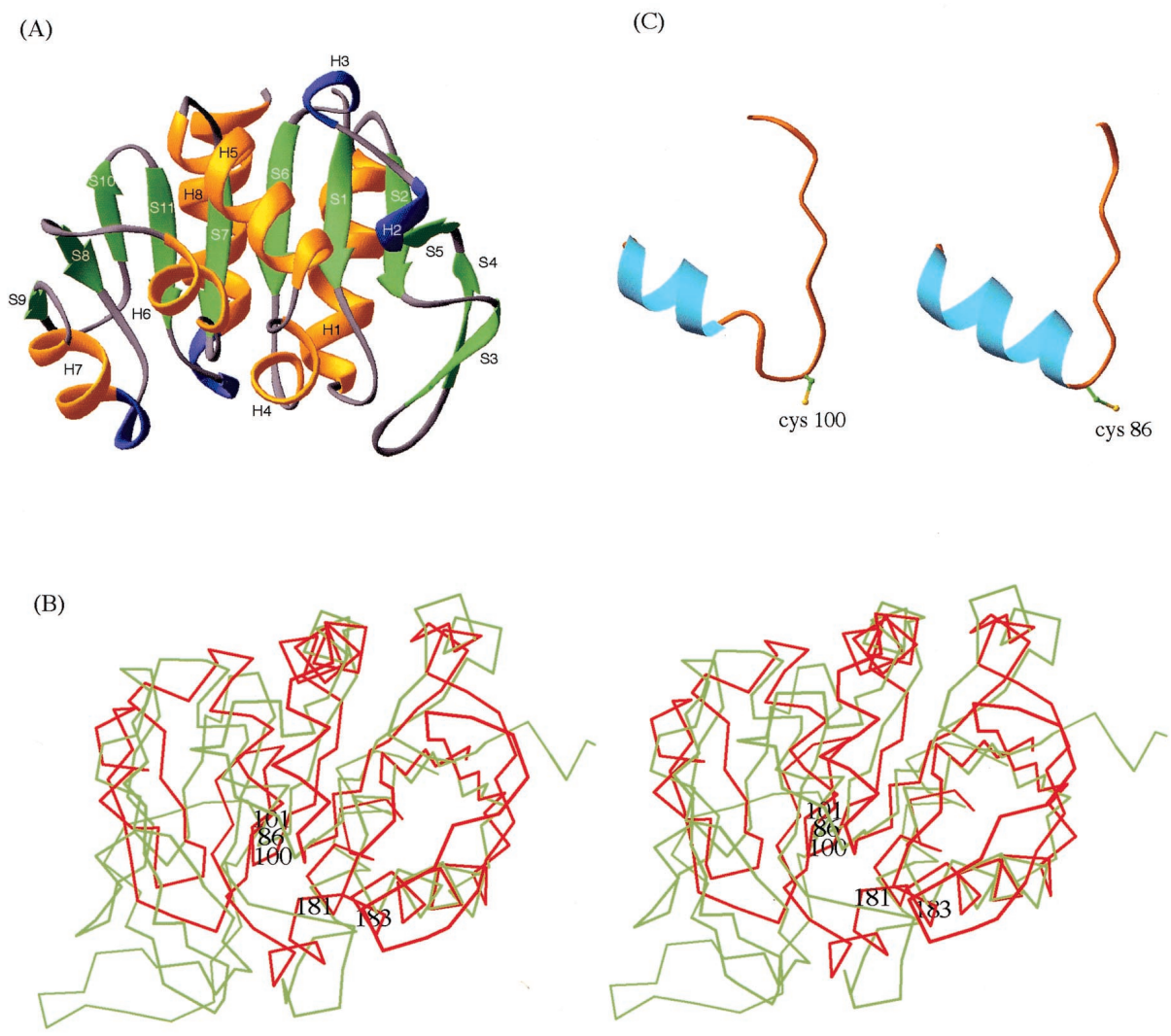

Fig. 3. Structural features of PH1704. (*A*) Ribbon diagram of PH1704. ''H'' and ''S'' refer to helix and strand, respectively. Green, gold, and blue code for β-strand, α-helix, and 3_{10} helix, respectively. (*B*) The superposition of the Cα trace of amidotransferase domain of GMP synthetase (green) and that of PH1704 (red). The three members of the catalytic triad, Cys-86, His-181, and Glu-183 of the amidotransferase and Cys-100 and His-101 of PH1704 are labeled. (*C*) The ''nucleophile elbows'' in PH1704 (*Left*) and amidotransferase domain of GMP synthetase (*Right*).

different mechanism. They all form barrel-like oligomeric structures with the active sites sequestered inside the barrel. The ATPase-containing regulatory complexes cover the outer port of the proteolytic chamber and regulate the entry of the substrates into the chamber. This feature of compartmentalization was also recognized in crystal structures of two ATP-independent intracellular proteases, Gal6 (29) and leucine aminopeptidase (LAP) (30). In each case, the active site resides in a cavity formed by a hexamer, and the access to the cavity is restricted by a small opening. Because the ATPase domain, which is thought to be involved in substrate unfolding and locomotion, is absent in these two structures, it was suggested that this class of compartmentalized proteases may specialize in the hydrolysis of small peptides that can freely permeate the small opening (31).

PH1704 and PfpI can be categorized into the same class as Gal6 and LAP. PfpI and PH1704 have been identified as ATP-independent proteases in a previous assay (2) and our gelatin-overlay protease assay described earlier, respectively. Although the overall structure of the PH1704 complex is more open, the active sites of PH1704 are located in hindered positions in the complex (Fig. 4*B*) and are not accessible to even the smallest globular protein. The active sites of PH1704 also lack the cleft that defined specificity and that binds to peptides of certain lengths, further suggesting that the protease may have a broad specificity. Interestingly, it was proposed that the physiological function of PfpI, a protein of high homology to PH1704, is to hydrolyze small peptides to provide a nutritional source for *P. furiosus* (3).

The presence of the nucleophile elbow in the PH1704 structure suggests that Cys-100 may be the active site nucleophile (Fig. 4*C*). In the crystal structure, Cys-100 is surrounded by Glu-12, Glu-15, Lys-43, His-101, Tyr-120, Val-150, and Pro-151 (Fig. 4*C*). Cys-100 also forms a ''catalytic triad'' with His-101 and Glu-474 (from an adjacent monomer). The catalytic triad in PH1704 shares the same handedness with the triads in papain type cysteine proteases, namely the cysteine interacts with δN, whereas the glutamate hydrogen-bonds with εN of the histidine (Fig. 4*D*). An aspect of the triad is its formation on a dimer interface, which is consistent with the experimental observation that the activity was found only for the oligomeric forms of the protein.

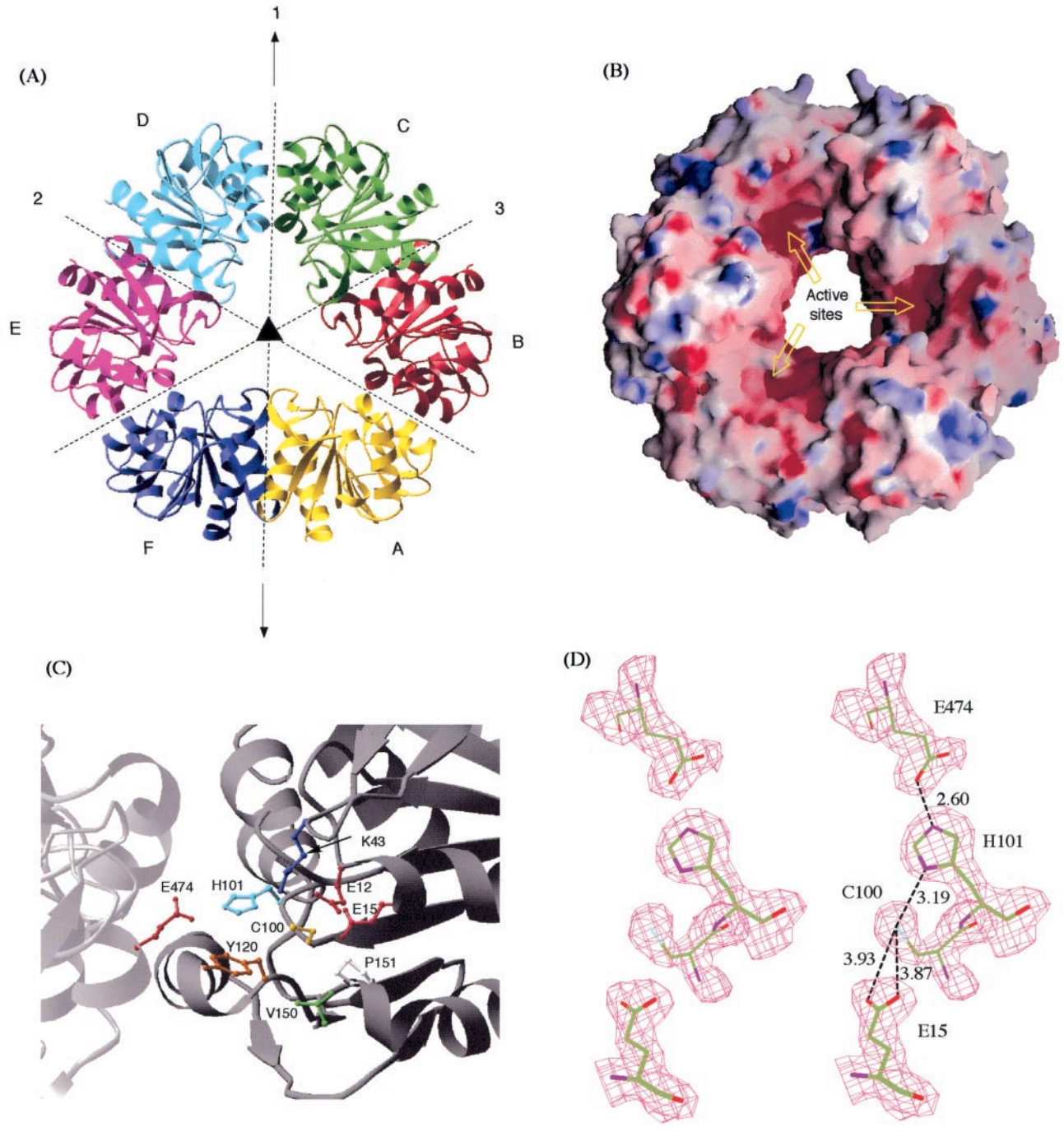

Fig. 4. Hexameric structures and putative catalytic sites of PH1704. (*A*) Ribbon diagram of the hexamer with the threefold NCS axis (perpendicular to the plane of paper) and the three twofold axes (in the plane of paper) are shown. Axis 1 coincides with a crystallographic symmetry axis. Axis 2 and 3 are NCS symmetry axes. (*B*) Surface representation of the hexamer color-coded by electrostatic potential. The diameter of the center opening is ≈24 Å. (*C*) Ball and stick model of the putative active site. All of the residues surrounding Cys-100 are shown. (*D*) A stereogram of the proposed catalytical triad and Glu-15 with the corresponding solvent-flattened experimental map. The distances shown are in Å.

We thank Dr. David King for characterizing PH1704 with electrospray mass spectrometry and Dr. Gerry McDermott for assistance in data collection. We also thank Dr. Edward Berry, Zhaolei Zhang, Dr. Alfonso Martinez, Dr. Dong-Hae Shin, and Dr. Luhua Lai for their help throughout the project. The work was supported by the Director, Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract no. DE-AC03-76SF00098.

1. Blumentals, I. I., Robinson, A. S. & Kelly, R. M. (1990) *Appl. Env. Microbiol.* **56,** 1992–1998.

2. Halio, S. B., Blumentals, I. I., Short, S. A., Merrill, B. M. & Kelly, R. M. (1996) *J. Bacteriol.* **178,** 2605–2612.

3. Snowden, L., Blumentals, I. I. & Kelly, R. (1992) *Appl. Env. Microbiol.* **58,** 1134–1141.
4. Altschul, S. F. Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
5. Zeng, G. (1998) *BioTechniques* **25,** 206–208.
6. Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990) *Methods Enzymol.* **185,** 60–89.
7. Leahy, D. J., Hendrickson, W. A., Aukhil, I. & Erickson, H. P. (1992) *Science* **258,** 987–991.
8. Kim, R., Sandler, S. J., Goldman, S., Yokota, H., Clark, A. J. & Kim, S.-H. (1998) *Biotech. Lett.* **20,** 207–210.
9. Jancarik, J. & Kim, S.-H. (1991) *J. Appl. Crystallogr.* **24,** 409–411.
10. Otwinowski, Z. & Minor, W. (1997) *Methods Enzymol.* **277,** 307–327.
11. Dodson, E. J., Winn, M. & Ralph, A. (1997) *Methods Enzymol.* **277,** 620–633.
12. Terwilliger, T. C. & Berendzen, J. (1999) *Acta Crystallogr. D* **55,** 849–861.
13. Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990) *EMBO J.* **9,** 1665–1672.
14. Jones, A. & Kleywegt, G. (1997) *Methods Enzymol.* **277,** 173–208.
15. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., *et al.* (1998) *Acta Crystallogr. D* **54,** 905–921.
16. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
17. Holm, L. & Sander, C. (1993) *J. Mol. Biol.* **233,** 123–138.
18. Bravo, J., Mate, M. J., Schneider, T., Switala, J., Wilson, K., Loewen, P. C. & Fita I. (1999) *Proteins* **34,** 155–166.
19. Tesmer, J. J., Klem, T. J., Deras, M. L., Davisson, V. J. & Smith, J. L. (1996) *Nat. Struct. Biol.* **3,** 74–86.
20. Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., *et al.* (1992) *Protein Eng.* **5,** 197–211.
21. Wang, K. K. (2000) *Trends Neurosci.* **23,** 20–26.
22. Thornberry, N. A. & Lazebnik, Y. (1998) *Science* **281**, 1312–1316.
23. Allen, P. M., Babbitt, B. P. & Unanue, E. R. (1987) *Immunol. Rev.* **98,** 171–187.
24. Puri, J. & Factorovich, Y. (1988) *J. Immunol.* **141,** 3313–3317.
25. Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H. D. & Huber, R. *(1997) Nature (London)* **386,** 463–471.
26. Löwe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. & Huber, R. (1995) *Science* **268,** 533–539.
27. Wang, J., Hartling, J. A. & Flanagan, J. M. (1997) *Cell* **91,** 447–456.
28. Rohrwild, M., Pfeifer, G., Santarius, U., Muller, S. A., Huang, H. C., Engel, A., Baumeister, W. & Goldberg, A. L. (1997) *Nat. Struct. Biol.* **4,** 133–139.
29. Joshua-Tor, L., Xu, H. E., Johnston, S. A. & Rees, D. C. (1995) *Science* **269,** 945–950.
30. Burley, S. K., David, P. R., Taylor, A. & Lipscomb, W. N. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 6878–6882.
31. Larsen, C. N. & Finley, D. (1997) *Cell* **91,** 431–434.